

UC San Diego

UC San Diego Previously Published Works

Title

Improving transparency and scientific rigor in academic publishing.

Permalink

<https://escholarship.org/uc/item/28v5j1vm>

Journal

Brain and behavior, 9(1)

ISSN

2162-3279

Authors

Prager, Eric M
Chambers, Karen E
Plotkin, Joshua L
et al.

Publication Date

2019

DOI

10.1002/brb3.1141

Peer reviewed

Improving transparency and scientific rigor in academic publishing

ABSTRACT

Progress in basic and clinical research is slowed when researchers fail to provide a complete and accurate report of how a study was designed, executed, and the results analyzed. Publishing rigorous scientific research involves a full description of the methods, materials, procedures, and outcomes. Investigators may fail to provide a complete description of how their study was designed and executed because they may not know how to accurately report the information or the mechanisms are not in place to facilitate transparent reporting. Here, we provide an overview of how authors can write manuscripts in a transparent and thorough manner. We introduce a set of reporting criteria that can be used for publishing, including recommendations on reporting the experimental design and statistical approaches. We also discuss how to accurately visualize the results and provide recommendations for peer reviewers to enhance rigor and transparency. Incorporating transparency practices into research manuscripts will significantly improve the reproducibility of the results by independent laboratories.

SIGNIFICANCE

Failure to replicate research findings often arises from errors in the experimental design and statistical approaches. By providing a full account of the experimental design, procedures, and statistical approaches, researchers can address the reproducibility crisis and improve the sustainability of research outcomes. In this piece, we discuss the key issues leading to irreproducibility and provide general approaches to improving transparency and rigor in reporting, which could assist in making research more reproducible.

1 | INTRODUCTION

Progress in basic and clinical research is strongly dependent upon asking important research questions, attempting to answer those questions with robust methods, and then communicating the findings. Persuading colleagues that scientific results are objectively obtained and valid involves a willingness to report accurate, robust, and

transparent descriptions of the methods, procedures, and outcomes, which will allow for the independent replication, or reproducibility, of those findings (see Box 1 for definitions).

Publishers have the responsibility of providing a platform for the exchange of scientific information, while at the same time, it is the responsibility of the authors, journal editors, and peer reviewers to ensure that the published manuscripts are accurate. While many editors and peer reviewers expect that research published in their journals should be potentially reproducible, there are no set procedures to empirically test whether a finding can be independently reproduced. What's more, other barriers to reproducing results exist, including the laboratory environment, apparatus and test protocols, and animal strain (Crabbe, Wahlsten, & Dudek, 1999). A major source of irreproducibility also includes substantial systematic error, which can occur while scientists are conducting the experiments or during statistical analyses (Goodman et al., 2016). Systematic error can occur for a variety of reasons, including lack of scientific skill (e.g., two people performing the same experiment may not have the same level of experience) or variability in subject populations or reagents (Capes-Davis & Neve, 2016). In addition, when a researcher has inadequate statistical knowledge or there are honest flaws in the experimental design and statistical output, the errors generated might inappropriately influence the interpretation of the results (Baker, 2016; Steen, Casadevall, & Fang, 2013).

Efforts to improve research transparency (and, subsequently, reproducibility) by funders, researchers, and publishers have led to the development of checklists and new author guidelines (see, e.g., Cell Press' [Structured Transparent Accessible Reporting \[STAR\]](#) Methods and the *Journal of Neuroscience Research (JNR)* [Transparent Science Questionnaire](#)). However, checklists often go unchecked or unenforced by the publishers, editors, and/or peer reviewers (Baker, Lidster, Sottomayor, & Amor, 2014) and compliance by the authors is not always wholehearted (M. Macleod personal communication). Publishers cannot always ensure that the results are reproducible, but they can help the authors to present a transparent account of their work, including providing full details of the experimental and statistical procedures and results. Transparent and rigorous accounts of how an experiment was performed, why the authors used specific statistical approaches, and what limitations arise from such work will allow the reviewers, editors, and subsequently readers to better judge the quality of the science.

This article is simultaneously published in *Journal of Neuroscience Research* (<https://doi.org/10.1002/jnr.24340>) and *Cancer Reports* (<https://doi.org/10.1002/cnr.2.1150>).

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2018 The Authors. *Brain and Behavior* published by Wiley Periodicals, Inc.

Box 1 Definitions

1. *Open Science*—the process of making the content and process of producing evidence and claims transparent and accessible to others (Munafo et al., 2017).
2. *Methods Reproducibility*—complete and transparent reporting of information required for another researcher to repeat protocols and methods (Goodman, Fanelli, & Ioannidis, 2016).
3. *Results reproducibility*—independent attempts to reproduce the same or nearly identical results with the same protocols under slightly different conditions.
4. *Rigor*—applying the scientific method in the strictest sense to ensure an unbiased experimental design, analysis, interpretation, and reporting of results.
5. *Transparency*—the process by which the methodology, including the experimental design, data collection, coding, analysis, and tools used in data analysis, is clearly visible to all readers.
6. *Randomization*—the random allocation of participants/subjects to different experimental conditions or the order of sample collection to minimize the possibility of subjective influence in the assignment of subjects or unmeasured variables that might influence the outcome.
7. *Blinding*—the investigator and study staff are unaware of the group to which the subject was allocated from study onset through data analysis.

In this commentary, we offer an update to basic approaches in reporting a thorough account of the experimental design and statistical approaches and provide an overview of data visualization techniques (Weissgerber, Garovic, Winham, Milic, & Prager, 2016). It is our hope, as publishers and editors, that these guidelines will help the authors adhere to specific reporting guidelines that promote rigor and transparency in scientific research, which will ensure an accurate and complete account throughout their experiments and discourage publication bias. This, in turn, will promote better, more reproducible science.

2 | BARRIERS TO REPRODUCIBILITY

Many factors can lead to irreproducibility of scientific results. Oftentimes, these trace back to flaws in the experimental design, statistical analyses (and a lack of understanding of fundamental statistical principles), including low statistical power or inadequate sample sizes, basic reporting of the information essential for laboratories to independently reproduce results (e.g., biological reagents and reference material), and selective reporting of data/results (e.g., p-hacking) (Baker, 2016; Forstmeier, Wagenmakers, & Parker, 2017; Freedman, Cockburn, & Simcoe, 2015). These factors and others might contribute to between 50% and 90% of

the published papers being irreproducible (Begley & Ellis, 2012; Glasziou et al., 2014; Hartshorne & Schachner, 2012; Kilkenny et al., 2009; Macleod et al., 2014, 2015; Moher & Altman, 2015; van der Worp & Macleod, 2011). Attempts to reproduce published results cost the United States approximately \$28B annually (Freedman et al., 2015; Freedman, Venugopalan, & Wisman, 2017), yet poor descriptions of the published studies lead to a majority of studies becoming non-replicable (Glasziou et al., 2014). The next subsections will break down some of the more common barriers to reproducibility.

2.1 | Neglecting the methods and materials section in manuscripts

The materials and methods section of the manuscript is an often neglected area. Journals and authors often limit the methods section to brief descriptions of the procedures or place more complete methods into supplemental materials, or for journals moving away from supplemental material, to online methods that are separate from the article; these are not often critically reviewed by referees and can go unread by the experimenters. Furthermore, reviewers might not be able to adequately review methods and tools and subsequently might fail to notice that key details are missing. This can lead to a lack of complete and transparent reporting of the information required for another researcher to repeat protocols and methods (Goodman et al., 2016). Similarly, journals requiring a subsection on statistical analyses rarely ask the authors to provide a full account of the statistical approaches, and the authors may also fail to include a full account of the statistical outputs in the results section. Without a rigorous description of the methods, materials, and statistical approaches, experimenters lack the necessary information to independently replicate or nearly replicate results with the same protocol under similar conditions (Goodman et al., 2016; Kilkenny et al., 2009).

2.2 | Aiming for novelty and impact

Current publication trends place emphasis on the pursuit of novelty and innovation (Cohen, 2017), which leads to a collection of reporting problems in how data were obtained (Forstmeier et al., 2017). At the most extreme, pressure to publish may lead individuals to rush their experiments, cut corners, make unintentional errors in statistical outputs, or overinterpret the findings (Alberts, Kirschner, Tilghman, & Varmus, 2014), which can lead to irreproducibility of the scientific findings.

To publish in "high impact" journals, scientists may resort to submitting only their most novel and impactful findings and avoid presenting nonsignificant or incremental findings (Cohen, 2017), though the latter also have important implications in driving scientific progress. The pressure to publish sensational findings has even led some "high impact" journals to state in their submission forms: "negative results are not accepted" (Matosin, Frank, Engel, Lum, & Newell, 2014). This emphasis might encourage scientists to pursue nonlinear lines of investigation in search of statistical significance (e.g.,

p-hacking), and may be one driver of scientific misconduct, including falsifying and fabricating data to increase its impact or statistical significance (Steen et al., 2013). At the very least, it leads researchers to omit nonsignificant or incremental findings leading to a bias in the literature, and reinforces the perception that negative findings carry a low priority for publication (Capuano, Coats, Scavone, Rossi, & Rosano, 2015; Dickersin, Min, & Meinert, 1992). This publication bias has led science reporters and the public to declare that it has become more difficult to trust scientific findings (Bosman, 2006; Laine, Goodman, Griswold, & Sox, 2007).

2.3 | Inadequate training in experimental design, manuscript writing, and reporting tools

Even with the most rigorous reporting guidelines and stringent publication standards, including the precise application of the scientific method to ensure robust and unbiased experimental design, methodology, analysis, interpretation, and reporting of the results (Rigor and Reproducibility, 2017), it is not guaranteed the authors will fully comply. Reporting guidelines cannot overcome poor training in experimental design and statistics, both of which may be responsible for many of the challenges leading to irreproducibility (Collins & Tabak, 2014; Weissgerber, Garovic, Milin-Lazovic, et al., 2016). Indeed, investigators all too often make errors in designing and performing their research, in selecting statistical tests, and in reporting the results (Steward, 2016; Yamada & Hall, 2015). The problem can be exacerbated by errors being passed down by the primary investigator to students, by reviewers not catching these mistakes, and by editors not having the expertise to catch specific errors. However, tools to reeducate scientists at all levels in the experimental design and to employ correct data visualization techniques (Weissgerber et al., 2017; Weissgerber, Garovic, Savic, Winham, & Milic, 2016) are available (see the National Institutes of Health education modules designed to train students or retrain scientists on the responsible conduct of research, <https://www.nih.gov/research-training/rigor-reproducibility/training> or the [National Postdoctoral Association's Responsible Conduct of Research Toolkit](#)). Moreover, many institutions have statistical consultation available to investigators, which should be used; *JNR* and *Brain and Behavior* both hired statistical editors to review the submitted manuscripts for statistical accuracy, and *Current Protocols in Neuroscience* recently released a statistical guide that provides general guidelines regarding when, how, and why certain improved statistical techniques might be used in neuroscience research (Wilcox & Rousselet, 2018; see also Motulsky, 2014). These tools help the authors improve statistical reporting in manuscripts and ensure that the correct approach was used, though statistical reviews may be limited by how much raw data are available.

In addition to the above tools, editorials and commentaries published in various journals attempt to help the authors improve the descriptions of their experimental procedures and results to ensure that the published research is transparently and accurately reported (Bravo et al., 2015; Collins, Reitsma, Altman, & Moons, 2015; Hooijmans, Leenaars, & Ritskes-Hoitinga, 2010; Kilkenny, Browne, Cuthill, Emerson, & Altman, 2010a; Landis et al., 2012; Shamseer et al., 2015).

Unfortunately, the authors often fail to incorporate these guidelines into their articles and most journals do not enforce or penalize the authors for not including specific criteria (Baker et al., 2014). Refining the steps necessary to ensure quality control during the peer review and publication processes is essential in order to improve transparency and scientific rigor. Adopting the approaches discussed below will better ensure that the experimental designs are accurate and deviations from that design are explained, with the ultimate goal of increasing the reproducibility of the published data. Journals and publishers should continue to provide detailed guidelines to help the authors during the submission process, but if researchers do not adopt a rigorous and transparent approach to scientific design and reporting from the onset of training, these requirements will continue to fall short.

In the following sections, we outline the key steps to improve transparency and scientific rigor that should be considered during the designing stages of experiments, not just before submission for publication. These requirements can be broadly broken down into (a) reporting criteria to ensure rigor and transparency; (b) transparent account of experimental design; (c) improving statistical rigor and transparency; and (d) peer review to enhance rigor and transparency. Encouraging specific descriptions and a full account of the study will ensure transparency and could improve reproducibility efforts. The next four sections will break down these components to elaborate on how each can improve transparency and rigor in scientific reporting.

3 | REPORTING CRITERIA TO ENSURE RIGOR AND TRANSPARENCY

The following points describe the key characteristics that must be included in any research design to assess the internal validity, reliability, and potential for reproducibility of scientific findings. Many of these recommendations have been discussed in various venues (e.g., ARRIVE guidelines; Freedman et al., 2017; Kilkenny et al., 2010a; Munafo et al., 2017; Weissgerber, Garovic, Winham, et al., 2016; Weissgerber, Milic, Winham, & Garovic, 2015), and some might only be appropriate to specific sciences. However, we feel that inclusion of these criteria, when applicable, into research manuscripts will improve rigor and transparency of the experimental design and statistical approaches.

3.1 | Appropriately describing the experimental subjects

The methods section of each published study begins with a description of the experimental unit; however, in many cases, the information provided falls short. The experimental units are the entity that is randomly and independently assigned to the treatment conditions (e.g., human subject, animal, litter, cage, fish tank, culture dish; Lazic, Clarke-Williams, & Munafo, 2018). The sample size is equal to the number of experimental units. In considering the sample size, one

must ensure that the experimental units are independently allocated to the experimental condition, the application of the condition is applied independently to the unit, and the experimental units do not influence one another (Lazic et al., 2018). A significant concern in cell biology is determining whether cells or sections, for example, can be considered an experimental unit. In cases where an animal is treated and subsequent testing occurs postmortem (e.g., immunohistochemistry or electrophysiology), then the histological sections, neurons per section, spines per neuron, tumor cells per section, etc., are all subsamples of the experimental unit, which is the animal, and should be considered an *n* of 1 (Galbraith, Daniel, & Vissel, 2010; Lazic et al., 2018). If data are not independent, one strategy is to analyze clustered data (e.g., convert the replicates from a single subject into a single summary statistic; Galbraith et al., 2010). Alternatively, there are also procedures to accurately model the true variability in data sets using modern statistical techniques (e.g., handling nested data such as cells/animals, littermates; Wilson, Sethi, Lein, & Keil, 2017). As Stanley Lazic so eloquently concluded in his recent paper (Lazic, 2018),

...a few simple alterations to a design or analysis can dramatically increase the information obtained without increasing the sample size. In the interest of minimizing animal usage and reducing waste in biomedical research (Ioannidis et al., 2014; Macleod et al., 2014), researchers should aim to maximize power by designing confirmatory experiments around key questions, use focused hypothesis tests, and avoid dichotomizing and nesting that ultimately reduce power and provide no other benefits.

An appropriately written section describing the experimental subjects must include a statement of ethical approval (Institutional Review Board approval for human research or Institutional Animal Care and Use Committee approval for animals), followed by the total number of participants involved in each experiment. The authors must also include a clear description of the inclusion and exclusion criteria, which should be prespecified prior to the start of the experiments. Reporting the number of experimental units (i.e., subjects, animals, cells) excluded as well as the reason for exclusion is necessary to prevent the researcher from introducing selection bias that favors positive outcomes and distorts true effects (Holman et al., 2016). Crucially, studies involving human subjects must not reveal individual identifying information but must contain a full description of the participants' demographics as variations in the demographics can lead to confounding variables if not appropriately controlled. When designing an experiment, one must also account for sex as a biological variable (see below). One should carefully review the extant literature to determine whether sex differences might be observed in the study, and if so, design and power the study to test for sex differences. Omitting this step could compromise the rigor of the study (Clayton, 2016, 2018).

3.2 | Randomization and blinding procedures

Choices made by investigators during the design and execution of experiments can introduce bias, which may result in the authors reporting false positives (Kilkenny et al., 2009; Kilkenny, Browne, Cuthill, Emerson, & Altman, 2010b; Landis et al., 2012). For example, when investigators are aware of which animals belong to one condition or know that a given treatment should have a specific effect, or human subjects become aware of the conditions they are in, the researchers and participants may inadvertently be biased toward specific findings or alterations in a specific behavior (Karanicolas, Farrokhyar, & Bhandari, 2010; Schulz & Grimes, 2002). To reduce bias in subject and outcome selection, the authors should report randomization and blinding procedures (Festing & Altman, 2002). Implementing and reporting randomization and blinding procedures are simple and can be followed using a basic guide (Karanicolas et al., 2010; Smith, Morrow, & Ross, 2015), but to reduce bias, it is essential to report the method of participant randomization to the various experimental groups as well as on random sample processing and collection of data (Kilkenny et al., 2010a; Landis et al., 2012). Moreover, investigators should report whether experimenters are blind to the allocation sequence and also, in animal studies, report whether controls are true littermates of the test group (Galbraith et al., 2010). Similarly, once the investigator is blind to the conditions, they should remain unaware of the group in which the subject is allocated and the assessment outcome (Landis et al., 2012). Blinding is not always possible. In these cases, procedures to standardize the interventions and outcomes should be implemented and reported so groups are treated as equally as possible. In addition, researchers should consider duplicate assessment outcomes to ensure objectivity (Karanicolas et al., 2010). Attention to reporting these details will reduce bias, will avoid mistaking batch effects for treatment effects, and will improve the transparency of how the research was conducted.

3.3 | Animal housing and husbandry

Many life science disciplines use animal models to test their hypotheses. Few studies provide detailed information regarding housing and husbandry, and those reports that contain the information typically do not provide any level of detail that could allow for others to follow similar housing procedures. When using animals, care should be taken to adequately describe the housing and husbandry conditions as these conditions could have profound implications on the experimental results (Prager, Bergstrom, Grunberg, & Johnson, 2011). At a minimum, the authors should introduce in the abstract the race, sex, species, cell lines, etc., so that the reader will be aware of the population/sample being studied. However, in the methods section, the authors should carefully describe all animal housing and husbandry procedures. For example, it is normally unclear whether animals were single- or group-housed, and in most journals, the age and/or weight of the animals are commonly omitted (Florez-Vargas et al., 2016). Other factors that are not commonly reported include

information on how the animals were transported from a breeder to the experimenter vivarium (see Good practices in the Transportation of Research Animals, 2006), vivarium temperature, humidity, day/night schedules, how often cages are cleaned, how often animals are handled, whether enrichment is provided in a cage, and cage sizes (Prager et al., 2011). Requiring a full description of housing and husbandry procedures will be essential to the rigor and transparency of the published studies and could help determine why some studies are not reproducible.

3.4 | Sex as a biological variable

Sex/gender plays an influential role in experimental outcomes. A common practice within research is that findings in one sex (usually males) are generalized to the other sex (usually females). Yet, research consistently demonstrates that sex differences are present across disciplines. For example, as evidence reveals in a recent issue of JNR (see [Sex Influences on Nervous System Function](#)), sex not only matters at the macroscopic level, where male and female brains have been found to differ in connectivity (Ingallhalikar et al., 2014), but at the microscopic level too (Jazin & Cahill, 2010). The National Institutes of Health as well as a number of funding agencies mandates the inclusion of sex as a biological variable, yet this mandate is not enforced by most journals. Starting at the study design, the authors must review whether the extant literature suggests that sex differences might be observed in the study, and if so, then design and power the study to test for sex differences. Otherwise, the rigor of the study could be compromised. When publishing the results, the authors must account for sex as a biological variable, whenever possible. At a minimum, the authors should state the sex of the subjects studied in the title and/or abstract of the manuscript. The rationale for choosing only one sex if a single-sex study is conducted should also be provided, though discussed as a limitation to the generalizability of the findings. Investigators must also justify excluding either males or females. The assumptions that females are more variable than males or that females must be tested across the estrous cycle are not appropriate as these are not major sources of variability (Beery, 2018). This policy is not a mandate to specifically investigate sex differences, but requires investigators to consider sex from the design of the research question through reporting the results (Clayton, 2016, 2018). In some instances, sex might not influence the outcomes (e.g., Fritz, Amrein, & Wolfer, 2017; Segarra, Modamio, Fernandez, & Marino, 2017), but balancing sex in animal and cellular models will distinctly inform the various levels of research (Clayton, 2016). More specific guidelines for applying the policy of considering sex as a biological variable are also available (Clayton, 2018; McCarthy, Woolley, & Arnold, 2017), but shifting the experimental group composition should be done in the context of appropriate a priori power analyses. One concern is that sample sizes need to be doubled to identify effects using both female and male subjects, but factorial designs can evaluate the main effects of the treatment and subject sex without increasing the sample size (Collins, Dziek, Kugler, & Trall, 2014). While the risk of false-positive errors associated with

testing sex differences in this way is present, reporting that these differences may or may not be present is imperative to understanding how sex influences the function of the nervous system. This practice should be extended to all scientific journals using animal/human subjects.

3.5 | Transparent account of the experimental design and statistical approaches

A transparent experimental design, meaning how the experiment is planned to meet the specified objectives, describes all the factors that are to be tested in an experiment, including the order of testing and the experimental conditions. As studies become more complex and interconnected, planning the experimental procedures prior to the onset of experiments becomes essential. Yet even when the experiments are planned prior to their initiation, the experimental designs are often poorly described and rarely account for alterations in procedures that were used in the study under consideration. To provide a more transparent and rigorous approach to describing the experimental design, a new section should be placed after the "subjects" paragraph describing, in detail, the experimental design and deviations made from the original design.

The experimental design section should consist of two main components: (a) a list of the experimental procedures that were used to conduct the study, including the sequence and timing of manipulation; and (b) an open discussion of any deviations made from the original design. The description should include an explanation of the question(s) being tested, whether this is a parameter estimation, model comparison, exploratory study, etc.; the dependent and independent variables; replicates (how often the experiments were performed and how the data were nested); and the type of design considered (e.g., completely randomized design, randomized complete block design, and factorial design; see Lin, Zhu, & Su, 2015; Suresh, 2011) for definitions and procedures to implement these designs. Assuming the authors planned the analysis prior to data collection, the authors should describe the specific a priori consideration of the statistical methods and planned comparisons (Weissgerber, Garovic, Winham, et al., 2016) or report that no a priori statistical planning was carried out. If the statistical approach deviated from how it was originally designed (see, e.g., Registered Reports below), the authors should also report the justification for this change. This open description could help to improve independent research reproducibility efforts and assist reviewers and readers in understanding the rationale for specific approaches.

A precise description of how methodological tools and procedures are prepared and used should also be provided in the experimental design section. Oftentimes, methodological procedures are truncated, forcing the authors to omit critical steps. Alternatively, the authors may report that the methods were previously described but might have modified those procedures without reporting those changes. Due to current publishing constraints, various caveats that go into the methodological descriptions remain unknown. However, this can be remedied easily by journals requiring a full description or

step-by-step procedure of the experimental protocol used to test the dependent variables. Two options are available for publishing full protocols. First, the protocol could be published in the manuscript, with the reviewers verifying that the procedures are appropriately followed; second, a truncated version of the methods could be published in the manuscript, but the extended methods must be required as supplemental material (the extended methods will be peer-reviewed during the submission process). An alternative approach is to deposit step-by-step protocols into a database or a data repository such as Dryad, Figshare, or with the Center for Open Science, where they will receive a DOI and can be linked back to the original research article, which will contain the truncated procedures.

3.5.1 | Materials

Rigorous descriptions of the experimental protocols not only require a level of detail in the description of the experimental design, but also require a full account of the resources and how they were prepared and used. A contributing factor to irreproducibility is the poor or inaccurate description of materials. In order for researchers to replicate and build upon published research findings, they must have confidence in knowing that materials specified in a publication can be correctly identified so that they might obtain the same materials and/or find out more about those materials. Most studies do not include sufficient detail to uniquely identify key research resources, including model organisms, cell lines, and antibodies, to name a few (Vasilevsky et al., 2013). While most author guidelines request that the authors provide the company name, city in which the company is located, and the catalog number of the material, (a) many authors do not include this information; (b) the particular product may no longer be available; or (c) the catalog number or lot number is reported incorrectly, thus rendering the materials unattainable.

A new system is laying the foundation to report research resources with a unique identification number that can be deposited in a database for quick access. The Resource Identification Initiative standardizes the materials necessary to conduct research by assigning research resource identifiers (RRIDs; Bandrowski et al., 2016). To make it as simple as possible to obtain RRIDs, a platform was developed (www.scicrunch.org/resources) to aggregate data about antibodies, cell lines, model organisms, and software into a community database that is automatically updated on a weekly basis and provides the most recent articles that contain RRIDs. While SciCrunch is among the founding platforms, these identifiers can also be found on other sites, including antibodyregistry.org, benchsci.com, and others. Similarly, though more involved, PubChem offers identification for various compounds such as agonists and antagonists. Simply find the chemical abstract service (CAS) number from the chemical safety data sheet (SDS), input that number into PubChem, and receive the PubChem Chemical Identifier (CID). RRIDs have been successfully implemented in many titles throughout Wiley and are also in use by Cell Press and a number of other publishers. The authors should provide RRIDs and CIDs when describing resources such as antibodies, software (including statistical software used, as this is

rarely reported), and model organisms, or compounds used, allowing for easy verification by peer reviewers and experimenters.

3.5.2 | Statistical rigor and transparency

With most statistical software having a user-friendly interface, students quickly learn how to perform basic statistical tests. However, users all too often choose inadequate and incorrect statistical methods or approaches or cannot reproduce their analyses since they have only a rudimentary understanding to each test and when to use them (Baker et al., 2014; Lazic, 2010; Strasak, Zaman, Marinell, Pfeiffer, & Ulmer, 2007; Weissgerber, Garovic, Milin-Lazovic, et al., 2016). What's more, the authors do not appropriately describe their statistical approaches in text, partially because tests are performed only after the study is executed. In designing and reporting the experiments, the authors should report normalization procedures, tests for assumptions, exclusion criteria, and why statistical approaches might differ from what the authors originally proposed, if they developed these approaches prior to the onset of data collection. In addition, the authors must also include the statistical software and specific version thereof, descriptive statistics, and a full account of the statistical outputs in the results section.

Errors in statistical outputs often arise when the authors (a) do not conduct and report a power calculation (Strasak et al., 2007) or do not distinguish between exploratory and confirmatory analyses (Kimmelman, Mogil, & Dirnagl, 2014); (b) fail to state which statistical tests are used or provide adequate detail about the tests, including the descriptive statistics and a full account of the statistical output; (c) fail to state whether assumptions were examined (Weissgerber et al., 2015); or (d) fail to describe how replicates were analyzed (Lazic, 2010). Moreover, it might be difficult to reproduce statistical output when the authors do not report the statistical software and specific version thereof, fail to include in the manuscript the exclusion criteria or code used to generate analyses, or explain how modifications to the experimental design might lead to changes in how statistical analyses are approached (e.g., independent vs. non-independent groups; additional details about these common mistakes can be found in Refs: Weissgerber, Garovic, Milin-Lazovic, et al., 2016; Weissgerber, Garovic, Winham, et al., 2016; Weissgerber et al., 2017), but it is important to emphasize that failure to report these variables can lead to errors in data interpretation.

Choosing the correct statistical analyses first depends on an appropriate experimental design and mode of investigation (exploratory vs. confirmatory; Kimmelman et al., 2014). One must decide whether experimental conditions are independent, meaning that no subjects or specimens are related to each other (Weissgerber et al., 2017; Weissgerber, Garovic, Winham, et al., 2016), whether the conditions are non-independent or paired, and whether there are any associations between variables (Nayak & Hazra, 2011). The second step is that statistical analyses must include specific details about the test statistics, rationale for choosing each test, a description of whether normal distribution patterns are obtained, and a statement about which p-value level is deemed statistically significant. In

addition, a transparent and rigorous statistical analysis section must include the following:

- Power analysis calculations or sample size justification for exploratory research, including accuracy in parameter estimation (Maxwell, Kelley, & Rausch, 2008).
- Statement of the factors tested, types of analyses, and what post hoc comparisons were made.
- Statement of the statistical tests used and details as to why those tests were chosen, including how the authors choose between parametric and nonparametric tests (assumptions aside).¹
- Statement of an assessment of assumptions.
- Statement of how replicates were analyzed. (e.g., are western blots performed in duplicate and band pixels averaged?)
- Data point exclusion criteria.
- Statement of how outliers were determined and how they were handled.
- Descriptions of raw data, including transformation procedures.
- Within the results, a full account of the test statistic, and where applicable, the degrees of freedom, p-values reported to a consistent number of decimal places (usually three), and statement of whether the test was one- or two-sided.

3.5.3 | Power analysis

Many studies are rejected for publication because of criticism that a study is underpowered, though many more studies are published despite this (Button et al., 2013). Reporting how a sample size was predetermined based on power analyses conducted during the experimental design stage is a good way to avoid this criticism. Researchers are taught to perform these analyses prior to the start of their experiments, but evidence suggests that researchers and peer reviewers do not fully understand the concept of statistical power, have not been given adequate education about the concept, or do not consider the measurement important in designing the experiments (Onwuegbuzie & Leech, 2004).

Reviewers and journal editors are beginning to ask authors to address the question of what the power of the study was to detect the observed effect (Goodman & Berlin, 1994; Levine & Ensom, 2001). Determining whether a study is appropriately powered a priori or post hoc is a matter of debate (Levine & Ensom, 2001). Many argue that post hoc power analyses are inappropriate, especially for non-significant findings, while others argue that post hoc power analyses are appropriate since a priori power analyses do not represent the power of the ensuring effect, but rather the hypothesized effect (Onwuegbuzie & Leech, 2004).

The a priori power analysis is the most common way of determining the sample size for simple experiments and can be easily

computed using freely available software such as [G*Power](#). The sample size depends on a mathematical relationship among the (a) effect size of interest; (b) standard deviation (SD); (c) chosen significance level; (d) chosen power; and (e) alternative hypothesis (Festing & Altman, 2002). Yet, as more parameters come into play (e.g., within mixed-effects modeling), power analysis software becomes more complex (see [Power Analysis for Mixed Effect Models in R](#)). Conducting these analyses allows researchers to confidently select a sample size large enough to lead to a rejection of the null hypothesis for a given effect size (Onwuegbuzie & Leech, 2004). However, one limitation to a priori power analyses is that effect sizes and SDs may not be known prior to the research being conducted and may lead to observed effects that are smaller or larger than the hypothesized effects (Wilkinson & Inference, 1999; see also Nuzzo, 2014). Alternatively, if it is conventional to use a specific number of subjects for a particular test, then one can report the calculated effect size for that particular sample size and decide whether more samples would be warranted. Either way, power and sample size calculations provide a single estimate, ignoring variability and uncertainty as such simulations are highly encouraged (see Lazic, 2016).

An alternative to the a priori power analysis is a post hoc power analysis (SPSS calls this “observed power”) or confidence intervals. The post hoc power analysis takes the observed effect size as the assumed population effect, though this computation might be different from a true population effect size, which might culminate in a misleading evaluation of power (Onwuegbuzie & Leech, 2004). Post hoc power analyses always show there is low power with respect to nonsignificant findings (Levine & Ensom, 2001). Thus, utilizing the post hoc power analysis must be done with extreme care and should never be a substitute for the a priori power analysis. In fact, many in the statistical community see post hoc analyses as a waste of effort and recommend abandoning this approach (Hoenig & Heisey, 2001); see also <https://dirnagl.com/2014/07/14/why-post-hoc-power-calculation-does-not-help/> and <https://daniellakens.blogspot.com/2014/12/observed-power-and-what-to-do-if-your.html>. If a reviewer or journal requests a power analysis, we recommend that rather than using post hoc power analyses, report confidence intervals to estimate the magnitude of effects that are consistent with the statistical data reported (Goodman & Berlin, 1994; Levine & Ensom, 2001; Smith & Bates, 1992). Alternatively, if increasing power is a necessity and/or sample sizes are already at their limits for financial or logistic reasons, one should consider alternative approaches, which are well described by Lazic; these include (a) using fewer factor values for continuous predictors; (b) having a more focused and specific hypothesis test; (c) not dichotomizing or binning continuous variables; (d) using a crossed or factorial design rather than a nested arrangement (Lazic, 2018).

We also advise authors to determine whether a parametric or nonparametric test is the most appropriate for the obtained data. Analogues to ordinary parametric tests (e.g., t test or ANOVA, etc.) can be performed even if data are skewed or have nonnormal distributions; multiple robust analytics are available for these circumstances (see Wilcox, 2013) as long as the sample size is sufficient.

When describing the data, it is important to differentiate between an exploratory and confirmatory study, as this could have profound implications as to how data are presented. Exploratory analyses are meant to identify patterns in the data without much emphasis on hypothesis testing, but most studies publish confirmatory experiments to test one or a few stated hypotheses.

Importantly, parametric tests also generally have somewhat more statistical power than nonparametric tests and are more likely to detect a significant effect if one exists. Alternatively, when one's data are better represented by the median, nonparametric tests may be more appropriate, especially when data are skewed enough that a mean might be strongly affected by the distribution tail, whereas the median estimates the center of the distribution. Nonparametric tests may also be more appropriate when the obtained sample size is small, as occurs in many fields where sample sizes average less than eight per group (Holman et al., 2016) or when the data obtained are ordinal, ranked, or there are outliers that cannot be removed (Frost, 2015). Beware, however, that meaningful nonparametric testing with sample sizes too low (e.g., $n < 5$) contains very little appreciable power to reveal an effect, if indeed one is present; difficulties due to violations of the underlying statistical assumptions of the particular test being used might be present. Bayesian analyses with small sample sizes are also possible, though estimates are highly sensitive to the specification of the prior distribution.

3.5.4 | Graphical representation of data

Figures illustrate the most important findings from a study by conveying information about the study design in addition to showing the data and statistical outputs (Weissgerber et al., 2017; Weissgerber, Garovic, Winham, et al., 2016). Simplistic representations to visualize the data are commonly used and are often inappropriate. For example, bar graphs are designed for categorical data; when used to display continuous data, bar graphs with error bars omit key information about the data distribution (see also Rousselet, Foxe, & Bolam, 2016). To change standard practices for presenting data, continuous data should be visualized by emphasizing the individual points; dot plots (e.g., univariate scatterplots) are strongly recommended for small samples, along with plots such as violin plots (or overlaid points on the plots) to provide far more informative views of the data distributions when samples are sufficiently large. Bar graphs should be reserved for categorical data only. Moreover, graphic data plots involving multiple groups are often shown as overlaid, but should be "jittered" across the x-axis so that each discrete data point can be visualized. The use of jittering means that when there are fewer unique combinations of data points than total observations, the totality of the data distribution is not obscured. By adopting these practices, readers will be better able to detect gross violations of the statistical assumptions and determine whether results would be different using alternate strategies (Weissgerber et al., 2015).

When plotting data, it is important to also report the variability of the data. Typically, this is expressed as the *SD* or standard error of the mean (*SEM*), but it is important to note that *SEM* does indicate variability (Motulsky, 2014). The *SD* is calculated as part of an estimate of the variability of the population from which the sample was drawn (Altman & Bland, 2005; Nagele, 2003). The *SEM*, on the other hand, describes the *SD* of the sample mean as an estimate of the accuracy of the population mean. In other words, the *SD* shows how many points within the sample differ from the sample mean, whereas the *SEM*

shows how close the sample mean is to the population mean (Nagele, 2003). The main function of *SEM* is to help construct confidence intervals, which are a range of values that take into account the true population value (usually an unknown), so that one can quantify the proximity of the experimental mean to the population mean (Barde & Barde, 2012). Yet, deriving confidence intervals around one's data (using *SD*) or the mean (using *SEM*) is premised on those data being normally distributed. Robust estimators are increasingly important as heteroskedasticity (having subpopulations with differing variability) is a frequent consequence of real-world measurement. Traditional data transformations are an attempt to cope with this phenomenon, but for many, such transformations may not actually serve to resolve anything and may add a layer of unnecessary complexity.

In determining which estimate of variability to depict graphically, it is important to remember that the *SD* is used when one wants to know how widely scattered measurements are or the variability within the sample, but if one is interested in the uncertainty around the estimate of the mean measurement or the proximity of the mean to the population mean, *SEM* is more appropriate (Nagele, 2003). When plotting data variability, it is important to consider that when *SEM* bars do not overlap, the viewer cannot be sure that the difference between the two means is statistically significant (see Motulsky, 2014). We also note that it is misleading to report *SDs* in the narrative and tables but plot *SEMs*. Furthermore, unless an author specifically wants to inform the reader about the precision of the study, *SD* should be reported as it quantifies variability within the sample (Altman & Bland, 2005; Barde & Barde, 2012; Nagele, 2003). Therefore, the optimal method to visualize data variability is to display the raw data, but if that makes the graph too difficult to read, instead show a box-whisker plot, frequency distribution, or the mean \pm *SD* (Motulsky, 2014).

3.5.5 | Inclusion of statistically significant and nonsignificant data

The probability that a scientific research article is published traditionally depends on the novelty or inferred impact of the conclusion, the size of the effect measured, and the statistical confidence in that result (Matosin et al., 2014; Scargle, 2000). The consequence of obtaining negative results can lead to a file-drawer effect; scientists ignore negative evidence that does not reach significance and intentionally or unintentionally select the subsets of data that show statistical significance as the outcomes of interest (Munafo et al., 2017). This publication bias skews scientific knowledge toward statistically significant or "positive" results, meaning that the results of thousands of experiments that fail to confirm a result are filed away (Scargle, 2000). These data-contingent analysis decisions, also known as p-hacking (Simmons, Nelson, & Simonsohn, 2011), can inflate spurious findings and lead to misestimates that might have consequences for public health. To combat the stigma of reporting negative results, we encourage authors to provide a full account of the experiment, to explicitly state both statistically significant and nonsignificant results, and to publish papers that have been rigorously designed and conducted, irrespective of their statistical outcomes. In addition, some organizations such as the

European College of Neuropsychopharmacology are offering prizes in neuroscience research to encourage publication of data where the results do not confirm the expected outcome or original hypothesis (see [ECNP Preclinical Network Data Prize](#)). Published reports of both significant and nonsignificant findings will result in better scientific communication among and between colleagues.

3.5.6 | Real and perceived conflicts of interest

Though objectivity of a researcher or group is assumed, conflicts of interest may exist and could be a potential source of bias. Conflicts of interest largely focus on financial conflicts (Als-Nielsen, Chen, Gluud, & Kjaergard, 2003; Thompson, 1993), but they can also occur when an individual's personal interests are in conflict with professional obligations, including industrial relationships (Young, 2009). Conflicts, whether real or perceived, arise when one recognizes an interest as influencing an author's objectivity. This can occur when an author owns a patent, or has stock ownership, or is a member of a company, for example. All participants in a paper must disclose all relationships that could be viewed as presenting a real or perceived conflict of interest. When considering whether a conflict is present, one should ask whether a reasonable reader could feel misled or deceived. While beyond the scope of this article, the Committee on Publication Ethics offers a number of resources on [conflicts of interest](#).

3.5.7 | Registered reports and open practices badges

One possible way to incorporate all the information listed above and to combat the stigma against papers that report nonsignificant findings is through the implementation of Registered Reports or rewarding transparent research practices. Registered Reports are empirical articles designed to eliminate publication bias and incentivize best scientific practice. Registered Reports are a form of empirical article in which the methods and the proposed analyses are preregistered and reviewed prior to research being conducted. This format is designed to minimize bias, while also allowing complete flexibility to conduct exploratory (unregistered) analyses and report serendipitous findings. The cornerstone of the Registered Reports format is that the authors submit as a Stage 1 manuscript an introduction, complete and transparent methods, and the results of any pilot experiments (where applicable) that motivate the research proposal, written in the future tense. These proposals will include a description of the key research question and background literature, hypotheses, experimental design and procedures, analysis pipeline, a statistical power analysis, and full description of the planned comparisons. Submissions, which are reviewed by editors, peer reviewers, and in some journals, statistical editors, meeting the rigorous and transparent requirements for conducting the research proposed are offered an in-principle acceptance, meaning that the journal guarantees publication if the authors conduct the experiment in accordance with their approved protocol. Many journals often publish the Stage 1 report, which could be beneficial not only for citations,

but also for the authors' progress reports and tenure packages. Following data collection, the authors prepare and resubmit a Stage 2 manuscript that includes the introduction and methods from the original submission plus their obtained results and discussion. The manuscript will undergo full review; referees will consider whether the data test the authors' proposed hypotheses by satisfying the approved outcome-neutral conditions, will ensure the authors adhered precisely to the registered experimental procedures, and will review any unregistered post hoc analyses added by the authors to confirm they are justified, methodologically sound, and informative. At this stage, the authors must also share their data (see also Wiley's [Data Sharing and Citation Policy](#)) and analysis scripts on a public and freely accessible archive such as Figshare and Dryad or at the Open Science Framework. Additional details, including template reviewer and author guidelines, can be found by clicking the link to the [Open Science Framework](#) from the Center for Open Science (see also Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014).

The authors who practice transparent and rigorous science should be recognized for this work. Funders can encourage and reward open practice in significant ways (see <https://wellcome.ac.uk/what-we-do/our-work/open-research>). One way journals can support this is to award badges to the authors in recognition of these open scientific practices. Badges certify that a particular practice was followed, but do not define good practice. As defined by the Open Science Framework, three badges can be earned. The Open Data badge is earned for making publicly available the digitally shareable data necessary to reproduce the reported results. These data must be accessible via an open-access repository and must be permanent (e.g., a registration on the [Open Science Framework](#) or an independent repository at www.re3data.org). The Open Materials badge is earned when the components of the research methodology needed to reproduce the reported procedure and analysis are made publicly available. The Preregistered badge is earned for having a preregistered design, whereas the Preregistered+Analysis Plan badge is earned for having both a preregistered research design and an analysis plan for the research; the authors must report results according to that plan. Additional information about the badges, including the necessary information to be awarded a badge, can be found by clicking this link to the [Open Science Framework](#) from the Center for Open Science.

4 | PEER REVIEW TO ENHANCE RIGOR AND TRANSPARENCY

The process of peer review is designed to evaluate the validity, quality, and originality of the articles for publication. Yet, peer reviewers are not immune to making mistakes. For example, several studies were conducted where major errors were inserted into papers. In these studies, no reviewer ever found all the errors and some reviewers did not spot any errors (Godlee, Gale, & Martyn, 1998; Schroter et al., 2004). While it is beyond the scope of this article to discuss many of the defects of peer

review (see Smith, 2006), it is important to note that the changes to the peer review process are ongoing (Tennant et al., 2017) and publishers are working to develop more formal training processes. However, to quickly improve rigor and transparency in scientific research, peer review should emphasize the design and execution of the experiment. We are not saying that reviewers should focus solely on the experimental design; it is important for reviewers to weigh in on the novel insights of a study and how study results may or may not contribute to the field. However, to help ensure the accuracy and the validity of a study, emphasis should first be on the experimental design. To assist the reviewers, the authors should submit as part of their manuscript a [Transparent Science Questionnaire \(TSQ\)](#), or something equivalent, which identifies where in the manuscript specific elements that could aid in reproducibility efforts are found. The reviewers use this form to verify that the authors have included the relevant information and ensure that the study was designed and executed objectively, ensuring the study's validity and reliability. Using this or similar forms will also help reviewers to find the relevant information necessary to ensure the appropriateness of the design, which can then allow them to focus on the experimental outcomes. Adopting forms such as the TSQ or using services such as those offered by Research Square could also speed up the peer review process and reduce the cost in time committed by unpaid reviewers (which, in 2008, was estimated to cost \$2.3 billion) (<https://scholarlykitchen.sspnet.org/2010/08/31/the-burden-of-peer-review/>).

A multistage review where different parties are concerned with different aspects of the review may be optimal. Because many errors in manuscripts are found in the statistical output, one stage of review should be a statistical review, whereby a statistical editor reviews the statistical analyses of the manuscript to ensure accuracy, but also verifies that the most appropriate statistical tests for that design were used. Upon completion, the editor will then make a decision as to whether the approach and execution are sufficient and are in line with the reported statistical output. By having experts focus on specific aspects of a research report, journal editors will become more confident that the research published is valid and of high quality and integrity.

5 | CONCLUSIONS

A challenge in science is for scientists to be open and transparent about the procedures used to obtain results. A major source of irreproducibility is substantial human error, which can occur while scientists are conducting the experiments or during data/statistical analysis. Groups are continuing to develop systems that help researchers cover every aspect of the experimental design (e.g., [EQIPD](#) or [XDA](#)), but education and awareness of the key elements in research design and analysis are essential to transparent and reproducible research. By incorporating the specific elements discussed in this document into research manuscripts, researchers

can reduce subjective bias, while actively improving methods' reproducibility, which will increase the likelihood of research reproducibility as the two are closely linked (Goodman et al., 2016). While variability in results is inevitable, ensuring that every salient aspect of a study is reported will help others understand the procedures involved and potential sources of errors during the experimentation process, which will ultimately lead to greater transparency in science.

ACKNOWLEDGMENTS

We would like to thank Dr. Larry Cahill, Dr. Stanley Lazic, Dr. Hermina Nedelescu, Dr. Tracey Weissgerber, and Dr. Cora Lee Wetherington for valuable comments to this manuscript. EMP and AB acknowledge the contribution of the discussions that took place during the meetings organized by the ECNP Network Preclinical Data Forum (<https://www.ecnp.eu/research-innovation/ECNP-networks/List-ECNP-Networks/Preclinical-Data-Forum.aspx>).

A preprint of this paper, which includes a roadmap to follow when preparing original research manuscripts and comments made during the review of the paper can be found at <https://osf.io/5cvqh/>.

CONFLICT OF INTEREST

Dr. David McArthur serves as JNR's paid statistical reviewer and has reviewed in that capacity for other journals, both Wiley and other publishers. Dr. Anita Bandrowski runs SciCrunch, a company devoted to ensuring RRIIDs persist in the literature. Dr. Maryann Martone is a founder and the CSO of SciCrunch, which provides services supporting RRIIDs and is the Editor-in-Chief of Brain and Behavior. Dr. Eric Prager is the Editor-in-Chief of Journal of Neuroscience Research. Dr. Nidhi Bansal is the Editor-in-Chief of Cancer Reports. Chris Graf works for Wiley, and volunteers for COPE, Committee on Publication Ethics.

AUTHOR CONTRIBUTION

Conceptualization: EMP and CG; Writing—Original Draft: EMP and KC; Writing—Review and Editing: EMP, KC, JKP, DLM, AB, NB, MM, HCB, AB, and CG; and Supervision: CG. All authors take responsibility for the integrity and the accuracy of this manuscript.

ORCID

Eric M. Prager  <https://orcid.org/0000-0002-3810-0985>

Karen E. Chambers  <https://orcid.org/0000-0002-4716-0875>

Joshua L. Plotkin  <https://orcid.org/0000-0001-6232-7613>

David L. McArthur  <https://orcid.org/0000-0003-3385-1314>

Anita E. Bandrowski  <https://orcid.org/0000-0002-5497-0243>


Nidhi Bansal  <https://orcid.org/0000-0001-6502-3930>

Maryann E. Martone  <https://orcid.org/0000-0002-8406-3871>

Hadley C. Bergstrom  <https://orcid.org/0000-0003-3677-1619>

Anton Bespalov  <https://orcid.org/0000-0003-3730-1395>


Chris Graf  <https://orcid.org/0000-0002-4699-4333>


Eric M. Prager¹ 

Karen E. Chambers¹ 

Joshua L. Plotkin² 

David L. McArthur³ 

Anita E. Bandrowski⁴ 

Nidhi Bansal¹ 

Maryann E. Martone⁴ 

Hadley C. Bergstrom⁵ 

Anton Bespalov^{6,7} 

Chris Graf⁸ 

¹John Wiley & Sons, Inc., Hoboken, New Jersey

²Department of Neurobiology and Behavior, Stony Brook University,
Stony Brook, New York

³Department of Neurosurgery, David Geffen School of Medicine at
UCLA, Los Angeles, California

⁴Center for Research in Biological Systems, University of California at
San Diego, San Diego, California

⁵Department of Psychological Science, Program in Neuroscience and
Behavior, Vassar College, Poughkeepsie, New York

⁶Partnership for Assessment and Accreditation of Scientific Practice,
Heidelberg, Germany

⁷Valdman Institute of Pharmacology, Pavlov First State Medical
University, St. Petersburg, Russia

⁸John Wiley & Sons, Oxford, UK

Correspondence

Eric M. Prager, Journal of Neuroscience Research, John Wiley & Sons,
Inc., St. Hoboken, NJ.

Email: epgrafer@wiley.com

REFERENCES

- Alberts, B., Kirschner, M. W., Tilghman, S., & Varmus, H. (2014). Rescuing US biomedical research from its systemic flaws. *Proceedings of the National Academy of Sciences*, 111(16), 5773–5777. <https://doi.org/10.1073/pnas.1404402111>
- Als-Nielsen, B., Chen, W., Glud, C., & Kjaergard, L. L. (2003). Association of funding and conclusions in randomized drug trials: A reflection of treatment effect or adverse events. *JAMA*, 290(7), 921–928. <https://doi.org/10.1001/jama.290.7.921>
- Altman, D. G., & Bland, J. M. (2005). Standard deviations and standard errors. *BMJ*, 331(7521), 903. <https://doi.org/10.1136/bmj.331.7521.903>
- Baker, M. (2016). Is there a reproducibility crisis? *Nature*, 533, 452–454.
- Baker, D., Lidster, K., Sottomayor, A., & Amor, S. (2014). Two years later: Journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. *PLoS Biology*, 12(1), e1001756. <https://doi.org/10.1371/journal.pbio.1001756>
- Bandrowski, A., Brush, M., Grethe, J. S., Haendel, M. A., Kennedy, D. N., Hill, S., ... Initiative, R. R. I. (2016). The Resource Identification Initiative: A cultural shift in publishing. *Brain and Behavior*, 6(1), e00417. <https://doi.org/10.1002/brb3.417>
- Barde, M., & Barde, P. (2012). What to use to express the variability of data: Standard deviation or standard error of the mean? *Perspectives in Clinical Research*, 3(3), 113–116.
- Beery, A. K. (2018). Inclusion of females does not increase variability in rodent research studies. *Current Opinion in Behavioral Sciences*, 23, 143–149. <https://doi.org/10.1016/j.cobeha.2018.06.016>
- Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391), 531–533. <https://doi.org/10.1038/483531a>
- Bosman, J. (2006). Reporters find science journals harder to trust, but not easy to verify. *New York Times*.
- Bravo, E., Calzolari, A., De Castro, P., Mabile, L., Napolitani, F., Rossi, A. M., & Cambon-Thomsen, A. (2015). Developing a guideline to standardize the citation of bioresources in journal articles (CoBRA). *BMC Medicine*, 13, 33. <https://doi.org/10.1186/s12916-015-0266-y>
- Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365–376. <https://doi.org/10.1038/nrn3475>
- Capes-Davis, A., & Neve, R. M. (2016). Authentication: A standard problem or a problem of standards? *PLoS Biology*, 14(6), e1002477. <https://doi.org/10.1371/journal.pbio.1002477>
- Capuano, A., Coats, A. J., Scavone, C., Rossi, F., & Rosano, G. M. (2015). Disclosure of negative trial results. A call for action. *International Journal of Cardiology*, 198, 47–48. <https://doi.org/10.1016/j.ijcard.2015.06.157>
- Chambers, C., Feredoes, E., Muthukumaraswamy, S., & Etchells, P. (2014). Instead of “playing the game” it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*, 1, 4–17. <https://doi.org/10.3934/Neuroscience2014.1.4>
- Clayton, J. A. (2016). Studying both sexes: A guiding principle for biomedicine. *The FASEB Journal*, 30(2), 519–524. <https://doi.org/10.1096/fj.15-279554>
- Clayton, J. A. (2018). Applying the new SABV (sex as a biological variable) policy to research and clinical care. *Physiology & Behavior*, 187, 2–5. <https://doi.org/10.1016/j.physbeh.2017.08.012>
- Cohen, B. A. (2017). How should novelty be valued in science? *eLife*, 6, e28699. <https://doi.org/10.7554/elife.28699>
- Collins, L. M., Dziek, J. J., Kugler, K. C., & Trall, J. B. (2014). Factorial experiments. *American Journal of Preventive Medicine*, 47, 498–504. <https://doi.org/10.1016/j.amepre.2014.06.021>
- Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. (2015). Transparent reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). *Annals of Internal Medicine*, 162(10), 735–736. <https://doi.org/10.7326/L15-5093-2>
- Collins, F. S., & Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature*, 505(7485), 612–613.
- Crabbe, J. C., Wahlsten, D., & Dudek, B. C. (1999). Genetics of mouse behavior: Interactions with laboratory environment. *Science*, 284(5420), 1670–1672.
- Dickersin, K., Min, Y. I., & Meinert, C. L. (1992). Factors influencing publication of research results. Follow-up of applications submitted to two institutional review boards. *JAMA*, 267(3), 374–378. <https://doi.org/10.1001/jama.1992.03480030052036>
- Festing, M. F., & Altman, D. G. (2002). Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR Journal*, 43(4), 244–258. <https://doi.org/10.1093/ilar.43.4.244>
- Florez-Vargas, O., Brass, A., Karystianis, G., Bramhall, M., Stevens, R., Cruickshank, S., & Nenadic, G. (2016). Bias in the reporting of sex and age in biomedical research on mouse models. *eLife*, 5, e13615. <https://doi.org/10.7554/elife.13615>
- Forstmeier, W., Wagenmakers, E. J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings—A practical guide. *Biological Reviews of the Cambridge Philosophical Society*, 92(4), 1941–1968. <https://doi.org/10.1111/brv.12315>
- Freedman, L. P., Cockburn, I. M., & Simcoe, T. S. (2015). The economics of reproducibility in preclinical research. *PLoS Biology*, 13(6), e1002165. <https://doi.org/10.1371/journal.pbio.1002165>

- Freedman, L. P., Venugopalan, G., & Wisman, R. (2017). Reproducibility 2020: Progress and priorities. *F1000Research*, 6, 604. <https://doi.org/10.12688/f1000research.11334.1>
- Fritz, A. K., Amrein, I., & Wolfer, D. P. (2017). Similar reliability and equivalent performance of female and male mice in the open field and water-maze place navigation task. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, 175(3), 380–391. <https://doi.org/10.1002/ajmg.c.31565>
- Frost, J. (2015). *Choosing between a nonparametric test and a parametric test*. Retrieved from <https://blog.minitab.com/blog/adventures-in-statistics-2/choosing-between-a-nonparametric-test-and-a-parametric-test>
- Galbraith, S., Daniel, J. A., & Vissel, B. (2010). A study of clustered data and approaches to its analysis. *Journal of Neuroscience*, 30(32), 10601–10608. <https://doi.org/10.1523/JNEUROSCI.0362-10.2010>
- Glasziou, P., Altman, D. G., Bossuyt, P., Boutron, I., Clarke, M., Julious, S., ... Wager, E. (2014). Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet*, 383(9913), 267–276. [https://doi.org/10.1016/S0140-6736\(13\)62228-X](https://doi.org/10.1016/S0140-6736(13)62228-X)
- Godlee, F., Gale, C. R., & Martyn, C. N. (1998). Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: A randomized controlled trial. *JAMA*, 280(3), 237–240. <https://doi.org/10.1001/jama.280.3.237>
- Goodman, S., & Berlin, J. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine*, 121, 200–206. <https://doi.org/10.7326/0003-4819-121-3-199408010-00008>
- Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science Translational Medicine*, 8(341), 341ps312. <https://doi.org/10.1126/scitranslmed.aaf5027>
- Hartshorne, J. K., & Schachner, A. (2012). Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Neuroscience*, 6, 8. <https://doi.org/10.3389/fncom.2012.00008>
- Hoenig, J., & Heisey, D. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1), 1–6. <https://doi.org/10.1198/000313001300339897>
- Holman, C., Piper, S. K., Grittner, U., Diamantaras, A. A., Kimmelman, J., Siegerink, B., & Dirnagl, U. (2016). Where have all the rodents gone? The effects of attrition in experimental research on cancer and stroke. *PLoS Biology*, 14(1), e1002331. <https://doi.org/10.1371/journal.pbio.1002331>
- Hooijmans, C. R., Leenaars, M., & Ritskes-Hoitinga, M. (2010). A gold standard publication checklist to improve the quality of animal studies, to fully integrate the Three Rs, and to make systematic reviews more feasible. *Alternatives to Laboratory Animals*, 38(2), 167–182.
- Ingalhalikar, M., Smith, A., Parker, D., Satterthwaite, T. D., Elliott, M. A., Ruparel, K., ... Verma, R. (2014). Sex differences in the structural connectome of the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, 111(2), 823–828. <https://doi.org/10.1073/pnas.1316909110>
- Ioannidis, J. P., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., ... Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. *Lancet*, 383(9912), 166–175. [https://doi.org/10.1016/S0140-6736\(13\)62227-8](https://doi.org/10.1016/S0140-6736(13)62227-8)
- Jazin, E., & Cahill, L. (2010). Sex differences in molecular neuroscience: From fruit flies to humans. *Nature Reviews Neuroscience*, 11(1), 9–17. <https://doi.org/10.1038/nrn2754>
- Karanicolas, P. J., Farrokhyar, F., & Bhandari, M. (2010). Practical tips for surgical research: Blinding: Who, what, when, why, how? *Canadian Journal of Surgery*, 53(5), 345–348.
- Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., & Altman, D. G. (2010a). Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. *PLoS Biology*, 8(6), e1000412. <https://doi.org/10.1371/journal.pbio.1000412>
- Kilkenny, C., Browne, W., Cuthill, I. C., Emerson, M., & Altman, D. G. (2010b). Animal research: Reporting in vivo experiments: The ARRIVE guidelines. *British Journal of Pharmacology*, 160(7), 1577–1579. <https://doi.org/10.1111/j.1476-5381.2010.00872.x>
- Kilkenny, C., Parsons, N., Kadysewski, E., Festing, M. F., Cuthill, I. C., Fry, D., ... Altman, D. G. (2009). Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS ONE*, 4(11), e7824. <https://doi.org/10.1371/journal.pone.0007824>
- Kimmelman, J., Mogil, J. S., & Dirnagl, U. (2014). Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS Biology*, 12(5), e1001863. <https://doi.org/10.1371/journal.pbio.1001863>
- Laine, C., Goodman, S. N., Griswold, M. E., & Sox, H. C. (2007). Reproducible research: Moving toward research the public can really trust. *Annals of Internal Medicine*, 146(6), 450–453. <https://doi.org/10.7326/0003-4819-146-6-200703200-00154>
- Landis, S. C., Amara, S. G., Asadullah, K., Austin, C. P., Blumenstein, R., Bradley, E. W., ... Silberberg, S. D. (2012). A call for transparent reporting to optimize the predictive value of preclinical research. *Nature*, 490(7419), 187–191. <https://doi.org/10.1038/nature11556>
- Lazic, S. E. (2010). The problem of pseudoreplication in neuroscientific studies: Is it affecting your analysis? *BMC Neuroscience*, 11, 5. <https://doi.org/10.1186/1471-2202-11-5>
- Lazic, S. E. (2016). *Experimental design for laboratory biologists: Maximising information and improving reproducibility*. Cambridge, UK: Cambridge University Press.
- Lazic, S. E. (2018). Four simple ways to increase power without increasing the sample size. *Laboratory Animals*, 23677218767478. <https://doi.org/10.1177/0023677218767478>
- Lazic, S. E., Clarke-Williams, C. J., & Munafo, M. R. (2018). What exactly is 'N' in cell culture and animal experiments? *PLoS Biology*, 16(4), e2005282. <https://doi.org/10.1371/journal.pbio.2005282>
- Levine, M., & Ensom, M. H. (2001). Post hoc power analysis: An idea whose time has passed? *Pharmacotherapy*, 21(4), 405–409. <https://doi.org/10.1592/phco.21.5.405.34503>
- Lin, Y., Zhu, M., & Su, Z. (2015). The pursuit of balance: An overview of covariate-adaptive randomization techniques in clinical trials. *Contemporary Clinical Trials*, 45(Pt A), 21–25. <https://doi.org/10.1016/j.cct.2015.07.011>
- Macleod, M. R., Lawson McLean, A., Kyriakopoulou, A., Serghiou, S., de Wilde, A., Sherratt, N., ... Sena, E. S. (2015). Risk of bias in reports of in vivo research: A focus for improvement. *PLoS Biology*, 13(10), e1002273. <https://doi.org/10.1371/journal.pbio.1002273>
- Macleod, M. R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J. P., ... Glasziou, P. (2014). Biomedical research: Increasing value, reducing waste. *Lancet*, 383(9912), 101–104. [https://doi.org/10.1016/S0140-6736\(13\)62329-6](https://doi.org/10.1016/S0140-6736(13)62329-6)
- Matosin, N., Frank, E., Engel, M., Lum, J. S., & Newell, K. A. (2014). Negativity towards negative results: A discussion of the disconnect between scientific worth and scientific culture. *Disease Models & Mechanisms*, 7(2), 171–173. <https://doi.org/10.1242/dmm.015123>
- Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, 59, 537–563. <https://doi.org/10.1146/annurev.psych.59.103006.093735>
- McCarthy, M. M., Woolley, C. S., & Arnold, A. P. (2017). Incorporating sex as a biological variable in neuroscience: What do we gain? *Nature Reviews Neuroscience*, 18(12), 707–708. <https://doi.org/10.1038/nrn.2017.137>
- Moher, D., & Altman, D. G. (2015). Four proposals to help improve the medical research literature. *PLoS Medicine*, 12(9), e1001864. <https://doi.org/10.1371/journal.pmed.1001864>
- Motulsky, H. J. (2014). Common misconceptions about data analysis and statistics. *Journal of Pharmacology and Experimental Therapeutics*, 351(1), 200–205. <https://doi.org/10.1124/jpet.114.219170>
- Munafo, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., ... Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(0021), 1–9.

- Nagele, P. (2003). Misuse of standard error of the mean (SEM) when reporting variability of a sample. A critical evaluation of four anaesthesia journals. *British Journal of Anaesthesia*, 90(4), 514–516.
- Nayak, B., & Hazra, A. (2011). How to choose the right statistical test? *Indian Journal of Ophthalmology*, 59(2), 85–86. <https://doi.org/10.4103/0301-4738.77005>
- Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*, 506(7487), 150–152. <https://doi.org/10.1038/506150a>
- Onwuegbuzie, A., & Leech, N. (2004). Post hoc power: A concept whose time has come. *Understanding Statistics*, 3(4), 201–230. https://doi.org/10.1207/s15328031us0304_1
- Prager, E. M., Bergstrom, H. C., Grunberg, N. E., & Johnson, L. R. (2011). The importance of reporting housing and husbandry in rat research. *Frontiers in Behavioural Neurosciences*, 5, 38. <https://doi.org/10.3389/fnbeh.2011.00038>
- Rigor and Reproducibility. (2017). *Rigor and reproducibility*. Retrieved from <https://grants.nih.gov/reproducibility/index.htm>
- Rousselet, G. A., Foxe, J. J., & Bolam, J. P. (2016). A few simple steps to improve the description of group results in neuroscience. *European Journal of Neuroscience*, 44(9), 2647–2651. <https://doi.org/10.1111/ejn.13400>
- Scargle, J. (2000). Publication bias: The "file-drawer" problem in scientific inference. *Journal of Scientific Exploration*, 14(1), 91–106.
- Schroter, S., Black, N., Evans, S., Carpenter, J., Godlee, F., & Smith, R. (2004). Effects of training on quality of peer review: Randomised controlled trial. *BMJ*, 328(7441), 673. <https://doi.org/10.1136/bmj.38023.700775.AE>
- Schulz, K. F., & Grimes, D. A. (2002). Blinding in randomised trials: Hiding who got what. *Lancet*, 359(9307), 696–700. [https://doi.org/10.1016/S0140-6736\(02\)07816-9](https://doi.org/10.1016/S0140-6736(02)07816-9)
- Segarra, I., Modamio, P., Fernandez, C., & Marino, E. L. (2017). Sex-divergent clinical outcomes and precision medicine: An important new role for institutional review boards and research ethics committees. *Frontiers in Pharmacology*, 8, 488. <https://doi.org/10.3389/fphar.2017.00488>
- Shamseer, L., Moher, D., Clarke, M., Ghera, D., Liberati, A., Petticrew, M., ... the PRISMA-P Group (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. *BMJ*, 349, g7647. <https://doi.org/10.1136/bmj.g7647>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Smith, R. (2006). Peer review: A flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, 99(4), 178–182. <https://doi.org/10.1258/jrsm.99.4.178>
- Smith, A. H., & Bates, M. N. (1992). Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. *Epidemiology*, 3(5), 449–452. <https://doi.org/10.1097/00001648-199209000-00011>
- Smith, P., Morrow, R., & Ross, D. (2015). Randomization, blinding and coding. In P. Smith, R. Morrow, & D. Ross (Eds.), *Field trials of health interventions: A toolbox* (3rd ed., pp. 1–23). London, UK: Oxford University Press.
- Steen, R. G., Casadevall, A., & Fang, F. C. (2013). Why has the number of scientific retractions increased? *PLoS ONE*, 8(7), e68397. <https://doi.org/10.1371/journal.pone.0068397>
- Steward, O. (2016). A Rhumba of "r's": Replication, reproducibility, rigor, robustness: What does a failure to replicate mean? *eNeuro*, 3(4), 1–4. <https://doi.org/10.1523/ENEURO.0072-16.2016>
- Strasak, A. M., Zaman, Q., Marinell, G., Pfeiffer, K. P., & Ulmer, H. (2007). The use of statistics in medical research: A comparison of the New England Journal of Medicine and Nature Medicine. *The American Statistician*, 61(1), 47–55. <https://doi.org/10.1198/000313007X170242>
- Suresh, K. (2011). An overview of randomization techniques: An unbiased assessment of outcome in clinical research. *Journal of Human Reproductive Sciences*, 4(1), 8–11. <https://doi.org/10.4103/0974-1208.82352>
- Tennant, J. P., Dugan, J. M., Graziotin, D., Jacques, D. C., Waldner, F., Mietchen, D., ... Colomb, J. (2017). A multi-disciplinary perspective on emergent and future innovations in peer review. *F1000Research*, 6, 1151. <https://doi.org/10.12688/f1000research.12037.3>
- Thompson, D. (1993). Understanding financial conflicts of interest. *New England Journal of Medicine*, 329(8), 573–576. <https://doi.org/10.1056/NEJM199308193290812>
- van der Worp, H. B., & Macleod, M. R. (2011). Preclinical studies of human disease: Time to take methodological quality seriously. *Journal of Molecular and Cellular Cardiology*, 51(4), 449–450. <https://doi.org/10.1016/j.yjmcc.2011.04.008>
- Vasilevsky, N. A., Brush, M. H., Paddock, H., Ponting, L., Tripathy, S. J., Larocca, G. M., & Haendel, M. A. (2013). On the reproducibility of science: Unique identification of research resources in the biomedical literature. *PeerJ*, 1, e148. <https://doi.org/10.7717/peerj.148>
- Weissgerber, T. L., Garovic, V. D., Milin-Lazovic, J. S., Winham, S. J., Obradovic, Z., Trzeciakowski, J. P., & Milic, N. M. (2016). Reinventing biostatistics education for basic scientists. *PLoS Biology*, 14(4), e1002430. <https://doi.org/10.1371/journal.pbio.1002430>
- Weissgerber, T. L., Garovic, V. D., Savic, M., Winham, S. J., & Milic, N. M. (2016). From static to interactive: Transforming data visualization to improve transparency. *PLoS Biology*, 14(6), e1002484. <https://doi.org/10.1371/journal.pbio.1002484>
- Weissgerber, T. L., Garovic, V. D., Winham, S. J., Milic, N. M., & Prager, E. M. (2016). Transparent reporting for reproducible science. *Journal of Neuroscience Research*, 94(10), 859–864. <https://doi.org/10.1002/jnr.23785>
- Weissgerber, T. L., Milic, N. M., Winham, S. J., & Garovic, V. D. (2015). Beyond bar and line graphs: Time for a new data presentation paradigm. *PLoS Biology*, 13(4), e1002128. <https://doi.org/10.1371/journal.pbio.1002128>
- Weissgerber, T. L., Savic, M., Winham, S. J., Stanisavljevic, D., Garovic, V. D., & Milic, N. M. (2017). Data visualization, bar naked: A free tool for creating interactive graphics. *Journal of Biological Chemistry*, 292(50), 20592–20598. <https://doi.org/10.1074/jbc.RA117.000147>
- Wilcox, R. (2013). *Introduction to robust estimation and hypothesis testing*, Vol. 3. San Diego, CA: Academic Press.
- Wilcox, R. R., & Rousselet, G. A. (2018). A guide to robust statistical methods in neuroscience. *Current Protocols in Neuroscience*, 82, 8.42.41–8.42.30. <https://doi.org/10.1002/cpns.41>
- Wilkinson, L., & Inference, T. T. F. o. S. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604. <https://doi.org/10.1037/0003-066X.54.8.594>
- Wilson, M. D., Sethi, S., Lein, P. J., & Keil, K. P. (2017). Valid statistical approaches for analyzing sholl data: Mixed effects versus simple linear models. *Journal of Neuroscience Methods*, 279, 33–43. <https://doi.org/10.1016/j.jneumeth.2017.01.003>
- Yamada, K. M., & Hall, A. (2015). Reproducibility and cell biology. *The Journal of Cell Biology*, 209(2), 191–193. <https://doi.org/10.1083/jcb.201503036>
- Young, S. (2009). Bias in the research literature and conflict of interest: An issue for publishers, editors, reviewers and authors, and it is not just about money. *Journal of Psychiatry and Neuroscience*, 34, 412–417.